

Minimization Strategies

Random directions(RD): Compute random direction, then do Brent iterations in this direction. Only the computation of function value per iteration necessary.

Complexity: $O(1) \times B$ ($B = \#$ of Brent iterations).

Steepest Descent(SD): requires the gradient to do Brent iterations $\rightarrow n$ derivatives to be computed (Assumption: costs $O(1)$). Computation of the function and n derivatives results in $n + 1$ computations. Complexity $O(n + 1) + O(B)$.

Newton: computes the functional, the gradient and the Hessian. Complexity of function evaluation part: $O(n^2)$. Select direction, and then do Brent. Use either Gaussian elimination or Cholesky decomposition (recommended) to solve the system of n linear equations $\rightarrow O(n^3)$ matrix operations.

Minimization Strategies (cont.)

Spectral Method (SM): computes the second derivatives. Do Brent with the found direction and solve an eigenvector/eigenvalue-problem requiring $O(n^3)$ matrix operations and $O(n^2)$ function evaluations.

Overview on Convergence and Time Complexity:

	# fct. evaluations	matrix op.	convergence
Random directions	$O(1) \times O(B)$	–	linear
Steepest descent	$O(n) \times O(B)$	–	linear
Newton	$O(n^2) \times O(B)$	$O(n^3)$	quadratic
Spectral method	$O(n^2) \times O(B)$	$O(n^3)$	quadratic

Reminder: Saddle Points

Saddle point: point where the function has zero gradient, a minimum in some directions and a maximum in other directions.

Methods like Newton *cannot* differentiate between a minimum, a maximum or a saddle point. So in some cases they converge to saddle points instead of converging to a minimum.

(Remember remarks w.r.t. convex optimization! Terminal problem since Newton's method is applicable.)

In the special case of doing Newton with the Cholesky decomposition, we can detect that we are going into a saddle point, but we cannot avoid it. The spectral method will avoid saddle points automatically (the biggest difference to Newton).

The Spectral Method: Derivation

Goal: Avoid Saddle Points as solutions.

Strategy: **Approximate** the function by Taylor series:

$$f(\mathbf{x} + \Delta\mathbf{x}) = f(\mathbf{x}) + \Delta\mathbf{x} \cdot f'(\mathbf{x}) + \frac{1}{2} \Delta\mathbf{x} \cdot f''(\mathbf{x}) \cdot \Delta\mathbf{x}^T + O(\|\Delta\mathbf{x}\|^3)$$

and **perform an eigenvalue/-vector decomposition** of the symmetric matrix $f''(\mathbf{x}) = U\Lambda U^T$.

Assign $\mathbf{z} := \Delta\mathbf{x} \cdot U$ and $\mathbf{d} := U^T \cdot f'(\mathbf{x})$, where $UU^T = 1$ s.t.

$$\begin{aligned} f(\mathbf{x} + \Delta\mathbf{x}) &= f(\mathbf{x}) + \mathbf{z} \cdot \mathbf{d} + \frac{1}{2} \mathbf{z} \cdot \Lambda \cdot \mathbf{z}^T + O(\|\Delta\mathbf{x}\|^3) \\ &\approx f(\mathbf{x}) + \mathbf{z} \cdot \mathbf{d} + \frac{1}{2} \mathbf{z} \cdot \Lambda \cdot \mathbf{z}^T \end{aligned}$$

The Spectral Method (cont.)

We arrive at

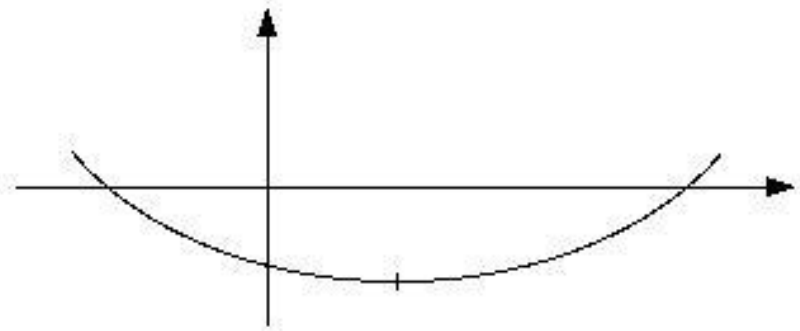
$$f(\mathbf{x} + \Delta\mathbf{x}) - f(\mathbf{x}) \approx \underbrace{\mathbf{z} \cdot \mathbf{d}}_{=\sum_i z_i d_i} + \frac{1}{2} \underbrace{\mathbf{z} \cdot \Lambda \cdot \mathbf{z}^T}_{=\sum_i z_i^2 \lambda_i}$$

The sum $\sum_i z_i d_i + \frac{1}{2} z_i^2 \lambda_i$ represents *n quadratic decoupled equations*.

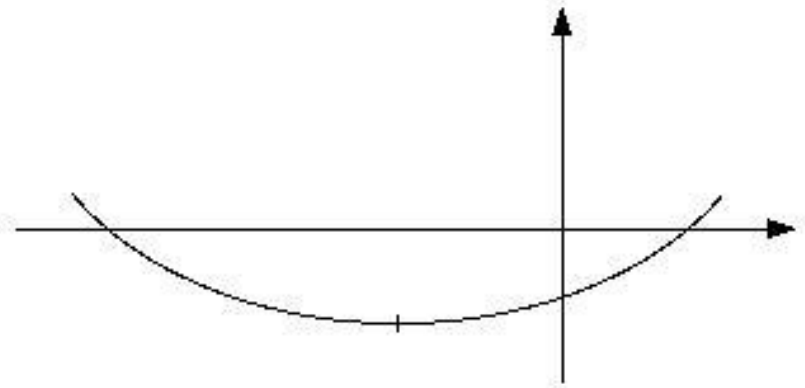
The z_i can be chosen to minimize each term *independently*, i.e. minimize $g_i(z_i) = z_i(d_i + \frac{1}{2} z_i \lambda_i)$ for each $1 \leq i \leq n$.

Depending on the value of the λ_i three cases can be distinguished.

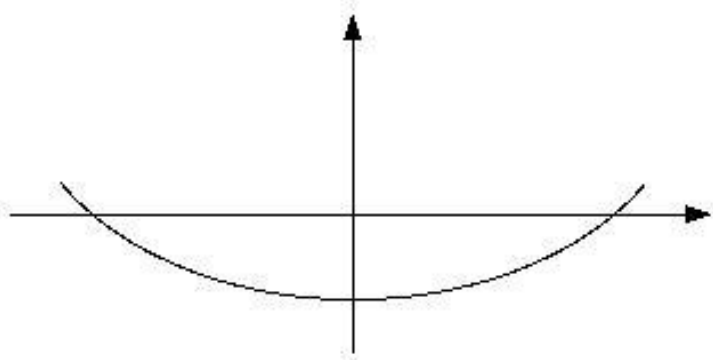
Case 1: $\lambda_i > 0$



$$d_i < 0: z_i = -\frac{d_i}{\lambda_i}$$

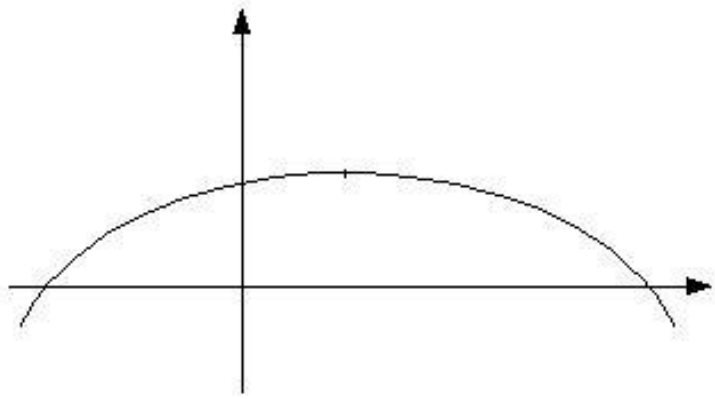


$$d_i > 0: z_i = -\frac{d_i}{\lambda_i} < 0$$



$d_i = 0$: Find minimum at $z_i = 0$.

Case 2: $\lambda_i < 0$



$$\lambda_i < 0$$

Maximum at $z_i = -\frac{d_i}{\lambda_i} > 0$.

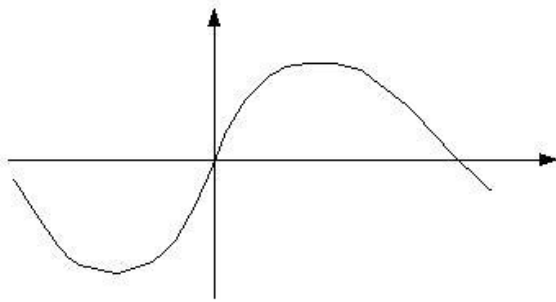
Note: approximation has no minimum. Choosing $|z_i| \gg 1 \rightsquigarrow$ parabola values arbitrarily small.

Newton, would lead us from $(0, 0)$ as starting point to the maximum \rightarrow *go in the opposite direction* (towards $-\text{sign}(-\frac{d_i}{\lambda_i})$).

How far?

Then our minimum would be at $z_i = -\text{sign}(-\frac{d_i}{\lambda_i})h$, where h is the unknown distance.

Case 2: $\lambda_i < 0$ (cont.)

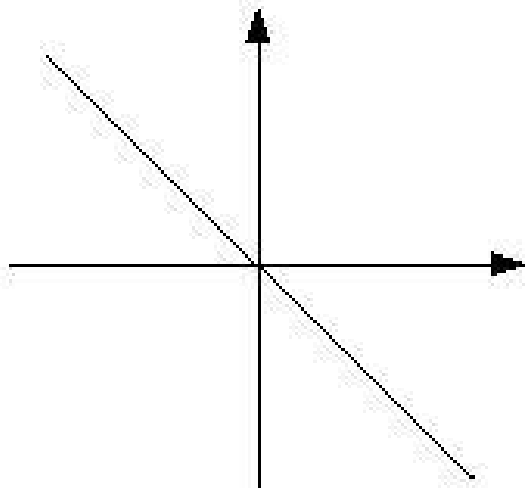


We ignored the $O(\|\Delta \mathbf{x}\|^3)$ -terms. The approximation only holds in a neighborhood of \mathbf{x} . \rightsquigarrow there can be other minima.

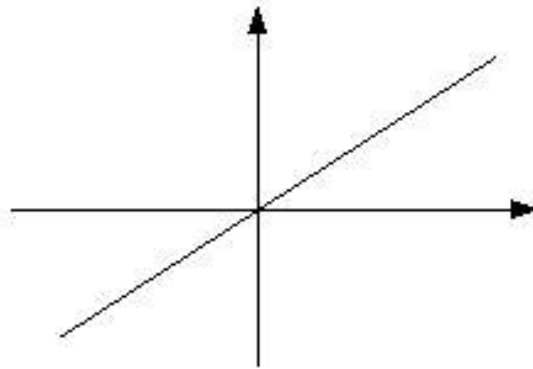
We choose the direction where we minimize the function with the closest z_i , yielding the best possible approximation.

For $d_i = 0$, (we are at the summit of the parabola) any direction has equal chance of success. Say, we choose “right”, i.e. if $\lambda_i < 0$, $d_i = 0$ then $z_i = h$.

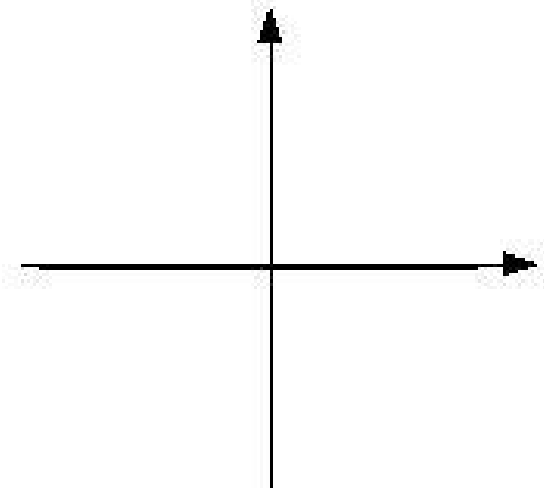
Case 3: $\lambda_i = 0$



$$d_i < 0:$$
$$z_i = -\text{sign}(d_i)h$$



$$d_i > 0:$$
$$z_i = -\text{sign}(d_i)h$$



$$d_i = 0:$$

best choice $z_i = 0$.

Comparison of minimization methods

Random Directions (RD) : cheapest method \Rightarrow use it as often as possible when it is useful.

RD is very effective at the beginning: we'll practically always find a direction which is not horizontal unless we are close to the minimum.

Steepest Descent (SD): second choice for the start of the minimization process, it too is inexpensive.

RD and SD share a lot of properties: very slow close to the end of the minimization process, i.e. near the minimum.

SD has a stairlike behaviour: Near the minimum the lines will be nearly at 90° angles. With higher order methods the approach is more direct.

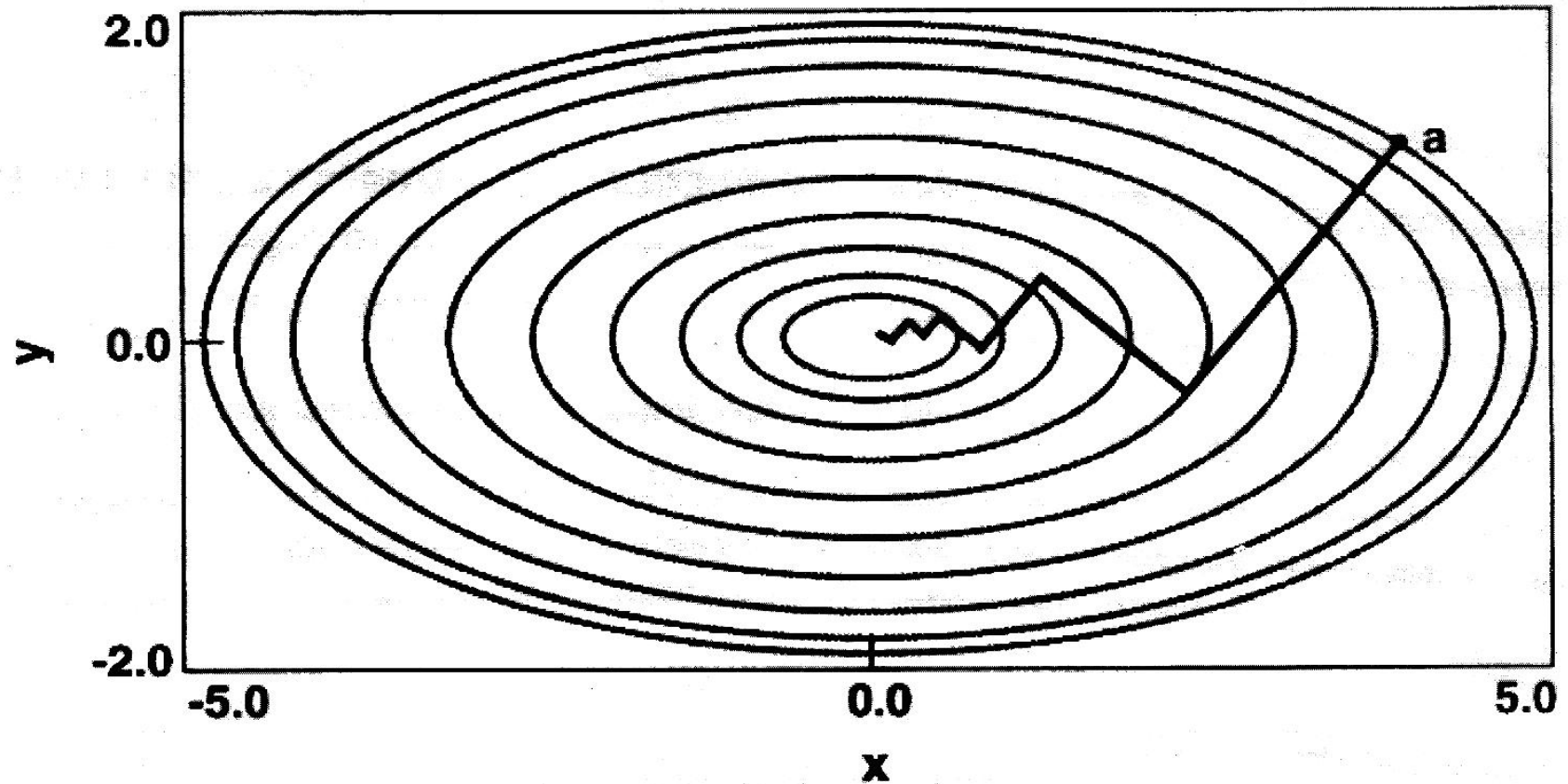


Figure 2–14. Minimization Path following a Steepest-Descent Path

Complete line searches starting from point **a** and converging on the minimum in about 12 iterations are used. In this case, where a rigorous line search is carried out, approximately 8 function evaluations are needed for each line search using a quadratic interpolation scheme. Note how steepest descents consistently overshoots the best path to the minimum, resulting in an inefficient oscillating trajectory.

Comparison of minimization methods (cont'd.)

Newton (N) and Spectral method (S) have quadratic convergence, i.e. they are very good, when you are close to the minimum. If the order of convergence is γ , then

$$\|x_{n+1} - x_n\| = O(\|x_n - x_{n-1}\|)^\gamma,$$

\Rightarrow with quadratic convergence, the number of accurate digits is basically doubled in each step.

Conclusion: RD and SD are cheap and simple \Rightarrow good in the beginning. N and SM are more expensive, but good in the end.

Comparison of minimization methods (cont'd)

N should always be used with **protection against saddle points**, i.e. a Cholesky decomposition:

- **Cholesky** decomposes a matrix A , appearing in an equation $Ax = b$ as $A = RR^T$, where R is lower triangular.
- **If A is not positive definite**, R cannot be computed, but at least we are prevented from converging to saddle points.
- If we use **Newton without Cholesky**, then **N** will converge for $n > 10$ almost inevitably to **saddlepoints instead of minima**.

When N fails due to a non positive definite A , we can either use SD (at the beginning) or S (at the end).

Comparison of minimization methods (cont'd.)

S is safest but most expensive.

In case that our functional is dependent on **fewer than n independent variables**, N will be forced to solve singular or (because of rounding errors) nearly singular problems and will **fail badly**.

S is robust in this sense: Tests for $\lambda_i = 0$ or $d_i = 0$ should take into account the roundoff error, i.e. test for $|\lambda_i| \leq \varepsilon$ and $|d_i| \leq \varepsilon$.

Conclusion

Recommended procedure: use RD (maybe SD) in the beginning. When there is not much progress, switch to a quadratic method (N with Cholesky), and whenever this fails use the spectral method.

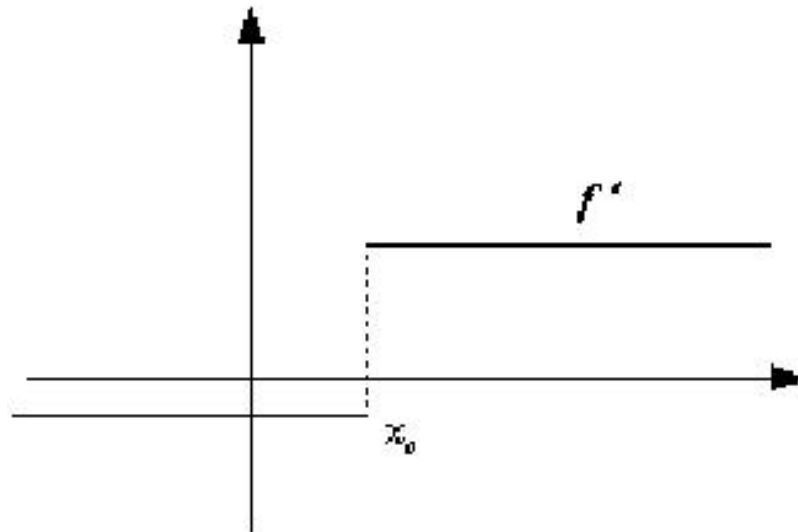
So we win the speed at the beginning (and get close to a local minimum, so that N has less chances of failing), and we **win the necessary speed in the end** with the quadratic methods.

Why should we code Newton *and* the Spectral method? S requires ~ 10 times more time than N \rightarrow severe problem for $n > 1000$ (in Molecular Dynamics, a small protein with 100 AAs already has $n = 900 \Rightarrow$ minimize number of eigenvalue/eigenvector calculations as much as possible.

Final Comments

Why use continuous functions with continuous derivatives?

Answer: Almost all methods rely on **Taylor expansion**: In N and SM the constants in the $O(\|\Delta^q \mathbf{x}\|^3)$ -term depend on the **third derivatives**. If third derivative is infinity at some point \Rightarrow error without bound!



Final Comments (cont'd)

Discontinuities in f, f' will (basically) **invalidate the approximation** \Rightarrow no quadratic convergence. Indeed often **no convergence at all!**.

When do such discontinuities arise? Unfortunately often because of oversights.

Example: structure of three atoms, C, O and S . Suppose, C and O are close together but S is far away.

C and the O close \Rightarrow **the vdW-forces relevant** . For distant S the vdW-forces are irrelevant (proportional to $O(\frac{1}{d^6})$.)

Example (cont'd)

Only compute the vdW-force inside a circle around atom-group:

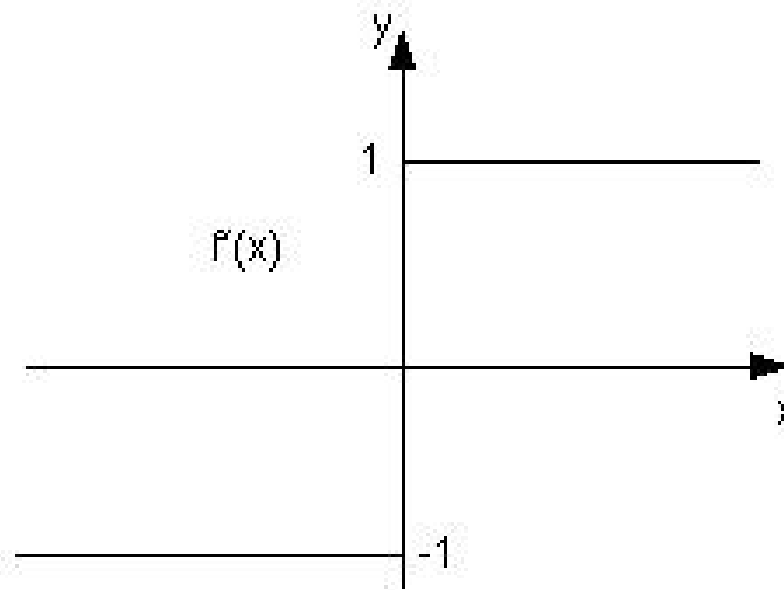
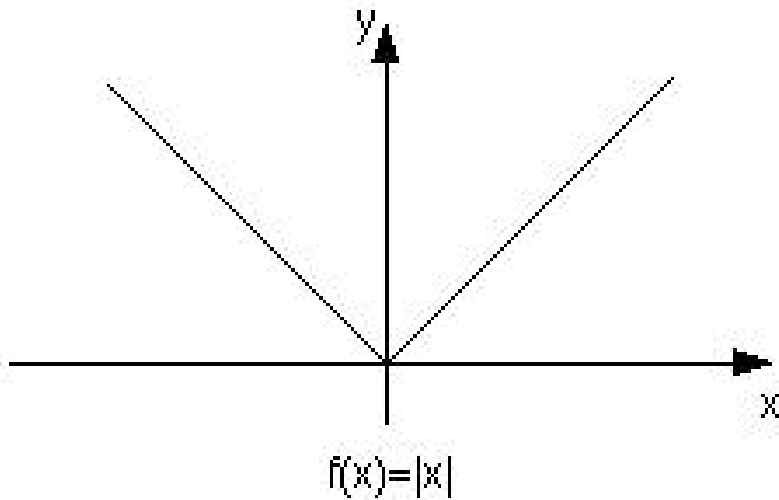
⇒ boundary: Inside we compute with the vdW-forces, outside we neglect them ⇒ discontinuity might invalidate the approximation.

It is important to make the functions continuous and differentiable!

Only if the contribution of a force will fall out of precision we can ignore a contribution.

Final Comments (cont'd)

Discontinuities also arise with the use of absolute value:
discontinuous first derivative:



Final Comments (cont'd)

The second final comment is about computational precision:
Should we use single precision, double precision or quadruple precision?

Possible answer: typically the distance between a carbon and a nitrogen atom this $d(C, N) = 1.8 \pm 0.05 \text{ \AA}$.
about 7 digits of precision.

Question: should we use more than single precision, when the initial data has merely two significant digits?

Answer: This question is seductive, but wrongly posed!

- When we say that the separation between the two atoms is 1.8 \AA , then we have a big error, i.e. we don't know this value with higher precision.

- But these two atoms and all other carbon and nitrogen atoms are separated by **exactly the same distance**. So for the minimization process this value should be considered as exact.
- The minimization process is very sensitive to errors (**approximation errors cause discontinuities!**). Working with 7 digits \Rightarrow some discontinuities will be introduced \Rightarrow approximations will become invalid.
- The precision that we have to use is **not related to the precision of the data**, it's related to the precision needed to run the **minimization process**.
- You will not be able to invert successfully a matrix of $100 \cdot 100$ or compute an eigenvalue/eigenvector decomposition in single precision.
- We need **at least double precision** to insure that we find a minimum, and that we don't stay stuck in some other place.